Online Micro-Gesture Recognition in Long Videos via Spatiotemporal Feature Encoding and Query-Based Temporal Detection

Chutian Meng, Fan Ma, Chi Zhang, Jiaxu Miao, Yi Yang and Yueting Zhuang*

Zhejiang University, 38 Zheda Road, Hangzhou 310027, China

Abstract

Micro-gestures (MGs) are subtle, involuntary body movements that can reveal underlying emotional or cognitive states, making them highly valuable for affective computing, psychological analysis, and human–computer interaction. However, recognizing MGs is particularly challenging due to their short duration, weak intensity, and lack of explicit boundaries. The MiGA2025 Challenge addresses these difficulties by providing a benchmark for micro-gesture recognition under realistic conditions, emphasizing cross-subject generalization and temporal localization. In this report, we present our solution for Track 2 of the MiGA2025 Challenge, which focuses on online micro-gesture recognition. To explore the broader potential of visual information, we conducted experiments on RGB data. We first employed the VideoMAE series of self-supervised video transformers to extract expressive spatiotemporal features from raw RGB inputs. These features were subsequently fed into the DyFADET network, a dynamic temporal modeling framework designed to detect fine-grained micro-gestures in continuous video streams. The results validate the effectiveness of combining pretrained visual representations with dynamic sequence modeling for advancing micro-gesture understanding in complex real-world settings.

Keywords

Micro-Gesture Detection, Temporal Action Detection, VideoMAE, DyFADet

1. Introduction

Micro-gestures (MGs) are subtle, involuntary human movements, often of the face, hands, or upper body, which can reveal hidden emotional states or intentions. Compared to overt gestures, MGs are short in duration, weak in intensity, and typically occur in unconstrained environments, making their automatic recognition a challenging task in the fields of affective computing, mental health analysis, and social signal processing.

To advance this area, the **MiGA Challenge** was introduced as part of IJCAI 2025. MiGA2025 offers three tracks focusing on micro-gesture understanding from skeleton and video modalities, with Track 2 specifically emphasizing **online micro-gesture recognition**. The task requires models to detect and classify spontaneous gestures in continuous video streams under cross-subject evaluation settings. This presents significant challenges including gesture sparsity, temporal ambiguity, and strong inter-subject variation.

In this work, we propose a pipeline that explores the use of RGB video to investigate how visual representations can enhance micro-gesture recognition. Compared to skeleton data, RGB input contains richer contextual and visual detail, including subtle texture changes and fine body movements, which may be crucial for recognizing micro-gestures. To tackle the challenges of subtle motion, sparsity, and long video durations in online micro-gesture detection, we design a two-stage pipeline that first extracts semantically rich clip-level features via self-supervised video transformers (VideoMAE), followed by a query-based temporal detector (DyFADet) capable of dynamic duration modeling and precise boundary localization.

MiGA@IJCAI25: International IJCAI Workshop on 3rd Human Behavior Analysis for Emotion Understanding, August 29, 2025, Guangzhou, China.

*Corresponding author.

12321215@zju.edu.cn (C. Meng); mafan@zju.edu.cn (F. Ma); 12321216@zju.edu.cn (C. Zhang); jiaxumiao@zju.edu.cn (J. Miao); yangyics@zju.edu.cn (Y. Yang); yzhuang@zju.edu.cn (Y. Zhuang)

Our contributions are as follows:

- **We explore the integration of RGB-based visual features in the context of micro-gesture recognition**, to study the potential benefit of vision-based pretraining for fine-grained motion understanding.
- **We apply the DyFADET network to model fine temporal dynamics of micro-gestures**, demonstrating strong boundary sensitivity and robustness under cross-subject evaluation.
- **We conduct extensive experiments on the SMG dataset**, showing that our pipeline achieves competitive performance and ultimately ranks **2nd place** in the MiGA2025 Track 2 leaderboard.

2. Related Work

2.1. Temporal Action Detection (TAD)

Temporal Action Detection (TAD) aims to localize action segments in untrimmed videos. Traditional two-stage methods, such as BSN [1] and BMN [2], generate proposals via boundary confidence modeling and then classify them. More recent one-stage detectors, like ActionFormer [3] and TCANet [4], directly regress temporal segments using multi-scale temporal features. Transformer-based approaches have significantly improved detection performance. TadTR [5] and RTD-Net [6] introduce end-to-end frameworks inspired by DETR, removing anchor designs and employing learnable queries. The recently proposed DyFADet [7] uses dynamic kernel aggregation and a flexible head structure to adapt to action durations of varying lengths.

2.2. Video Feature Extraction

Strong video representation is fundamental to TAD. Early methods used 3D CNNs like I3D [8] and SlowFast [9] to capture spatiotemporal patterns. Transformers further advanced this field: TimeSformer [10] employs divided space-time attention, and MViT [11] introduces a hierarchical token pyramid for scalable video modeling. Self-supervised pretraining has emerged as a powerful technique. VideoMAE [12] masks video patches for reconstruction, achieving competitive performance with minimal supervision. InternVideo [13] combines masked modeling and video-language contrastive learning, providing a general-purpose video foundation model.

2.3. Action Detection Frameworks

Action detection frameworks integrate backbone features and detection heads. Anchor-free frameworks (e.g., ActionFormer [3], TadTR [5]) simplify segment prediction by leveraging transformer decoders or temporal pyramids. Other designs like TCANet [4] emphasize context aggregation, while DyFADet [7] adaptively fuses multi-scale features for more robust detection. Recent reviews [14, 15, 16] provide a comprehensive categorization of TAD (Temporal Action Detection) methods and video foundation models, highlighting trends in self-supervised learning, query-based detection, and efficient feature extraction. In particular, **query-based detection** refers to approaches that formulate action localization as a set prediction problem, where a fixed number of learnable query vectors are used to decode potential action instances. These methods, inspired by DETR [17], remove the need for predefined anchors or dense proposals, and instead allow the model to directly attend to salient temporal regions through content-aware attention. DyFADet is a representative example of this paradigm, as it employs a learnable set of temporal queries to dynamically aggregate context across multiple scales for precise gesture localization.

3. Method

3.1. Overall Framework

Micro-gesture detection in untrimmed videos is a challenging task due to two main factors: (1) the **extreme video length**, with each video lasting up to 15 minutes, and (2) the **subtle and sparse nature** of micro-gestures, which are brief and occupy only a small portion of the video timeline. In light of these challenges, we adopt a **two-stage framework** in Figure 1 rather than an end-to-end approach that directly predicts gesture segments from raw RGB frames. The reasons are as follows:

- **Efficiency and Scalability.** End-to-end modeling over full-length videos is computationally intensive and memory-inefficient, especially at high frame rates. A two-stage approach allows us to first compress the video into compact clip-level features, significantly reducing the temporal and spatial resolution required in the detection stage.
- **Improved Temporal Modeling.** Micro-gestures are short and often occur in bursts within long videos. Decoupling feature extraction from detection enables us to preserve fine-grained motion cues without sacrificing global temporal context. Specialized detection models can then focus solely on temporal localization over the feature sequence.
- **Modular Optimization.** The two-stage design decouples visual representation learning and gesture localization. This modularity allows us to leverage powerful pretrained video backbones (e.g., VideoMAE) in Stage 1 and apply state-of-the-art temporal detectors (e.g., DyFADet) in Stage 2, benefiting from the best of both domains.

In summary, the two-stage architecture offers a practical and effective trade-off between computational feasibility and detection accuracy, making it well-suited for long-video micro-gesture analysis.

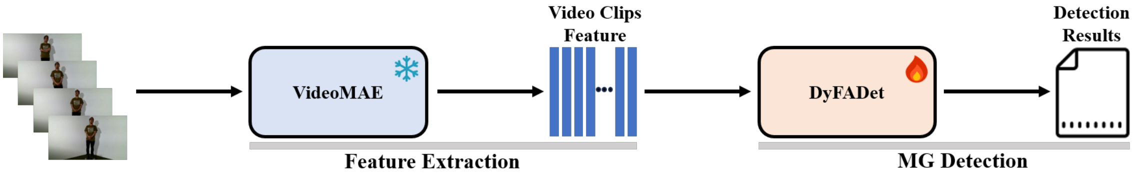

Figure 1: An overview of our two-stage micro-gesture recognition framework. In Stage 1, the long input video is segmented into frame clips and passed through VideoMAE for feature extraction. In Stage 2, the extracted features are fed into the DyFADet model to perform temporal localization and classification.

3.2. Stage 1: Feature Extraction

Given an untrimmed RGB video $V = \{f_1, f_2, \ldots, f_N\}$ with N frames, we divide it into a sequence of overlapping clips using a fixed sliding window. Following a strategy similar to THUMOS14, we extract clips every 4 frames, where each clip c_t consists of $L = 16$ consecutive frames. Each clip c_t is independently passed through a pretrained VideoMAE-g encoder $\phi(\cdot)$ to obtain a clip-level feature:

$$\mathbf{z}_t = \phi(c_t), \quad \mathbf{z}_t \in \mathbb{R}^d,$$

where is the feature dimension. This results in a compact temporal representation of the full video as a feature sequence:

$$\mathbf{Z} = \{\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_T\}, \quad \mathbf{Z} \in \mathbb{R}^{T \times d}.$$

We adopt VideoMAE [12] due to its strong spatiotemporal modeling capacity. VideoMAE is built on a Vision Transformer backbone and trained using a masked autoencoding objective. During pretraining, a large portion of video tokens is randomly masked, and the model learns to reconstruct the original input from the unmasked subset. This encourages the encoder to extract semantically rich and temporally aware representations. Compared to 3D CNNs or convolutional backbones, VideoMAE provides global receptive fields and better temporal context modeling, which is essential for capturing subtle micro-gesture motions.

In practice, we use the VideoMAE-g model pretrained on Kinetics-400 [18]. The encoder is frozen during downstream training, and only the gesture detection module (Stage 2) is trained. This separation enables efficient training and prevents overfitting, especially given the limited size of the target dataset.

3.3. Stage 2: Temporal Detection

Given the sequence of clip-level features $\mathbf{Z} = \{\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_T\}$ extracted from Stage 1, our goal is to detect temporal segments corresponding to micro-gestures. We formulate this as a temporal action detection task and adopt DyFADet [7], a recent anchor-free and query-based framework that achieves state-of-the-art performance in dense temporal localization.

Motivation. Compared with traditional proposal-based methods, DyFADet offers several advantages: (1) it avoids handcrafted anchor design and rigid grid constraints, (2) it leverages dynamic feature aggregation to adaptively model context at multiple temporal scales, and (3) it supports end-to-end training with a query-based detection head. [19]. These properties make DyFADet particularly suited for micro-gesture detection, where gestures may be subtle, sparse, and vary significantly in length.

Model Overview. DyFADet is temporal action detection model that takes a sequence of video features as input and produces a set of gesture predictions, each including start and end timestamps and a category distribution. A key component of DyFADet is its dynamic feature aggregation (DFA) module, which enables the model to focus adaptively on relevant temporal regions. Given a global feature sequence and a set of learnable queries, the DFA module computes localized context features via attention-based dynamic sampling. This mechanism allows DyFADet to model temporal structures more precisely and improves its ability to detect fine-grained micro-gestures under complex visual conditions.

Table 1
mAP@IoU and F1 Score on the SMG validation set. We compare different combinations of feature extractors (VideoMAE-b and VideoMAE-g) and detection heads (ActionFormer and DyFADet).

Method	Avg-mAP	mAP@0.3	mAP@0.4	mAP@0.5	mAP@0.6	mAP@0.7	F1 Score
VideoMAE-b + ActionFormer	9.62	14.13	12.36	10.97	6.84	5.25	0.3063
VideoMAE-b + DyFADet	10.59	15.22	13.78	11.65	7.50	5.46	0.3238
VideoMAE-g + ActionFormer	10.84	15.59	13.46	11.53	7.82	5.96	0.3306
VideoMAE-g + DyFADet (Ours)	**11.87**	**16.96**	**14.64**	**12.64**	**8.66**	**6.44**	**0.3634**

4. Experiments

4.1. Experimental Setup

Dataset. We conduct our experiments on the SMG dataset [20], which is designed for subtle micro-gesture understanding in long untrimmed videos. It consists of 40 videos, each approximately 15 minutes in duration, recorded in realistic environments. The given dataset includes 17 fine-grained gesture classes with manually annotated temporal boundaries. Following the official protocol, we divide the data into 30 videos for training, 5 for validation, and 5 for testing.

Evaluation Metrics. We adopt two types of evaluation metrics to comprehensively assess our model's performance. **(1) Temporal Localization Performance.** We report the mean Average Precision (mAP) at multiple temporal Intersection-over-Union (tIoU) thresholds: $\alpha \in \{0.3, 0.4, 0.5, 0.6, 0.7\}$.

A prediction is considered correct if it has the correct class label and a tIoU with a ground-truth segment exceeding the given threshold. The average of these scores is denoted as Avg-mAP. **(2) Online Recognition Performance.** Following the MiGA2025 Track2 protocol, we further evaluate the combined detection and classification performance using the F1 score:

$$\mathrm{F1} = \frac{2 \cdot \mathrm{Precision} \cdot \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}}.$$

Here, *Precision* is defined as the fraction of correctly classified micro-gestures (MGs) among all predicted gestures in the sequence, while *Recall* is the fraction of ground-truth MGs that are correctly retrieved by the algorithm. This metric jointly reflects temporal accuracy and class correctness, and is especially suitable for evaluating sparse and fine-grained actions in long untrimmed videos.

Implementation Details. We use the VideoMAE-g model pretrained on Kinetics-400. Kinetics-400 is a large-scale human action dataset containing approximately 240k training videos across 400 categories. It is widely used for pretraining video encoders due to its diversity and rich motion patterns, making it a suitable source for learning transferable spatiotemporal features. The DyFADet detection head is trained for 500 epochs using the AdamW optimizer with a learning rate of 1×10^{-4} and a batch size of 8. In the validation and testing phase, we set the number of prediction queries to 1000.All experiments are conducted on two NVIDIA A6000 GPUs.

4.2. Validation Results

We evaluate the effectiveness of our framework on the SMG validation set, consisting of 5 annotated long video sequences. As the test set ground-truth is unavailable, all quantitative results are reported on the validation set. We compare different combinations of video feature extractors and detection heads under various temporal IoU thresholds. Table 1 summarizes the performance. We observe that both the backbone and the detection head significantly influence final results:

1. Backbone impact: Using VideoMAE-g consistently outperforms VideoMAE-b across all detector settings. For example, when paired with ActionFormer, switching from VideoMAE-b to VideoMAE-g improves the average mAP from 9.62 to 10.84 and the F1 score from 0.3063 to 0.3306. This highlights the benefits of stronger spatiotemporal representations offered by larger transformer models. VideoMAE-g has approximately 304M parameters compared to 88M in VideoMAE-b, allowing it to better capture fine-grained motion cues that are essential for detecting subtle micro-gestures.

2. Detection head impact: DyFADet outperforms ActionFormer under both backbone configurations. Notably, the gains are more pronounced at higher IoU thresholds (e.g., mAP@0.7), suggesting DyFADet's superior boundary localization and flexible duration modeling. For instance, with VideoMAE-g, DyFADet achieves mAP@0.7 of 6.44 versus 5.96 from ActionFormer. This improvement can be attributed to DyFADet's query-based design and dynamic feature aggregation, which allow it to adaptively focus on relevant temporal segments without reliance on hand-crafted anchors.

3. Best configuration: Our final setup, VideoMAE-g combined with DyFADet, achieves the highest overall performance on validation set, with an average mAP of 11.87 and an F1 score of 0.3634. This confirms the effectiveness of our two-stage design and justifies the use of both a large-scale video backbone and a flexible query-based temporal detector for micro-gesture analysis.

5. Potential Improvements

While our proposed framework achieved strong performance in the MiGA2025 Challenge, several limitations remain that could be addressed in future work to further enhance robustness, efficiency, and interpretability.

1) Fine-tuning the backbone. Our current pipeline freezes the pretrained VideoMAE encoder during downstream training. While this prevents overfitting on limited micro-gesture data, it may lead to suboptimal alignment between visual features and detection objectives. Fine-tuning the backbone jointly with the detection head could allow task-specific gradients to refine the representation, potentially improving detection sensitivity and boundary precision.

2) Efficient long video processing. We employ a fixed sliding window strategy for dense clip extraction, which leads to high computational cost, especially for 15-minute videos. This could result in redundant or overlapping clips and wasted computation. Exploring adaptive clip sampling strategies based on motion saliency or confidence-driven attention could reduce redundancy and accelerate inference without significant performance loss.

3) Multi-modal integration. Our method exclusively relies on RGB input. While visual texture and motion cues are essential, micro-gestures often involve subtle joint movements that may be difficult to capture visually. Incorporating skeletal pose data (e.g., via pose estimation models such as OpenPose or MediaPipe), depth, or infrared modalities could provide complementary geometric cues, improving recognition in cases of low contrast, occlusion, or subtle movement.

4) Temporal imbalance and class skew. The SMG dataset contains highly imbalanced gesture distributions, with some categories appearing much more frequently than others. Our current model does not apply explicit balancing strategies during training. Techniques such as class-aware sampling, cost-sensitive loss functions, or dynamic reweighting could mitigate this imbalance and improve performance on underrepresented classes.

5) Post-hoc interpretability. Transformer-based components such as VideoMAE and DyFADet offer strong performance but often lack transparency. In sensitive applications such as mental health or education, trust and interpretability are essential. Post-hoc techniques such as attention rollout, temporal saliency maps, or class-specific activation tracing could provide insight into model behavior, failure cases, and potential biases.

6. Declaration on Generative AI

During the preparation of this work, the author(s) used ChatGPT, Grammarly in order to: Grammar and spelling check, Paraphrase and reword. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication‘s content.

7. Acknowledgments

This work is supported by the National Key R&D Program of China under Grant 2022ZD0160101 and National Natural Science Foundation of China (U2336212).

References

[1] T. Lin, X. Zhao, H. Su, C. Wang, M. Yang, Bsn: Boundary sensitive network for temporal action proposal generation, in: ECCV, 2018.

[2] T. Lin, X. Liu, X. Li, E. Ding, S. Wen, Bmn: Boundary-matching network for temporal action proposal generation, in: ICCV, 2019.

[3] C.-L. Zhang, J. Wu, Y. Li, Actionformer: Localizing moments of actions with transformers, in: European Conference on Computer Vision, volume 13664 of *LNCS*, 2022, pp. 492–510.

[4] Z. Qing, H. Su, W. Gan, D. Wang, W. Wu, X. Wang, Y. Qiao, J. Yan, C. Gao, N. Sang, Temporal context aggregation network for temporal action proposal refinement, in: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 485–494. doi:10.1109/CVPR46437.2021.00055.

[5] X. Liu, Q. Wang, Y. Hu, X. Tang, S. Zhang, S. Bai, X. Bai, End-to-end temporal action detection with transformer, IEEE Transactions on Image Processing 31 (2022) 5427–5441. doi:10.1109/TIP.2022.3195321.

[6] J. Tan, J. Tang, L. Wang, G. Wu, Relaxed transformer decoders for direct action proposal generation, in: 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 13506–13515. doi:10.1109/ICCV48922.2021.01327.

[7] L. Yang, Z. Zheng, Y. Han, H. Cheng, S. Song, G. Huang, F. Li, Dyfadet: Dynamic feature aggregation fornbsp;temporal action detection, Springer-Verlag, Berlin, Heidelberg, 2024, p. 305–322. URL: https://doi.org/10.1007/978-3-031-72952-2_18. doi:10.1007/978-3-031-72952-2_18.

[8] J. Carreira, A. Zisserman, Quo vadis, action recognition? a new model and the kinetics dataset, 2018. URL: https://arxiv.org/abs/1705.07750. arXiv:1705.07750.

[9] C. Feichtenhofer, H. Fan, J. Malik, K. He, Slowfast networks for video recognition, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019.

[10] G. Bertasius, H. Wang, L. Torresani, Is space-time attention all you need for video understanding?, in: ICML, 2021.

[11] H. Fan, B. Xiong, K. Mangalam, Y. Li, Z. Yan, J. Malik, C. Feichtenhofer, Multiscale vision transformers, 2021. URL: https://arxiv.org/abs/2104.11227. arXiv:2104.11227.

[12] Z. Tong, Y. Song, J. Wang, L. Wang, Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training, NeurIPS (2022).

[13] Y. Wang, K. Li, Y. Li, Y. He, B. Huang, Z. Zhao, H. Zhang, J. Xu, Y. Liu, Z. Wang, S. Xing, G. Chen, J. Pan, J. Yu, Y. Wang, L. Wang, Y. Qiao, Internvideo: General video foundation models via generative and discriminative learning, 2022. URL: https://arxiv.org/abs/2212.03191. arXiv:2212.03191.

[14] K. Hu, C. Shen, T. Wang, et al., Overview of temporal action detection based on deep learning, Artificial Intelligence Review (2024).

[15] M. C. Schiappa, Y. S. Rawat, M. Shah, Self-supervised learning for videos: A survey, ACM Comput. Surv. 55 (2023). URL: https://doi.org/10.1145/3577925. doi:10.1145/3577925.

[16] N. Madan, A. Moegelmose, R. Modi, Y. S. Rawat, T. B. Moeslund, Foundation models for video understanding: A survey, 2024. URL: https://arxiv.org/abs/2405.03770. arXiv:2405.03770.

[17] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, S. Zagoruyko, End-to-end object detection with transformers, 2020. URL: https://arxiv.org/abs/2005.12872. arXiv:2005.12872.

[18] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, A. Zisserman, The kinetics human action video dataset, 2017. URL: https://arxiv.org/abs/1705.06950. arXiv:1705.06950.

[19] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, S. Zagoruyko, End-to-end object detection with transformers, in: Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I, Springer-Verlag, Berlin, Heidelberg, 2020, p. 213–229. URL: https://doi.org/10.1007/978-3-030-58452-8_13. doi:10.1007/978-3-030-58452-8_13.

[20] H. Chen, H. Shi, X. Liu, X. Li, G. Zhao, Smg: A micro-gesture dataset towards spontaneous body gestures for emotional stress state analysis, International Journal of Computer Vision 131 (2023) 1346–1366.